Predicting Hazardous Driving Events Using Multi-Modal Deep Learning Based on Video Motion Profile and Kinematics Data

Z. Gao, Y. Liu, J. Y. Zheng, *Senior Member*, *IEEE*, R. Yu, X. Wang, and P. Sun

***Abstract*—As the raising of traffic accidents caused by commercial vehicle drivers, more regulations have been issued for improving their safety status. Driving record instruments are required to be installed on such vehicles in China. The obtained naturalistic driving data offer insight into the causal factors of hazardous events with the requirements to identify where hazardous events happen within large volumes of data. In this study, we develop a model based on a low-definition driving record instrument and the vehicle kinematic data for post-accident analysis by multi-modal deep learning method. With a higher camera position on commercial vehicles than cars that can observe further distance, motion profiles are extracted from driving video to capture the trajectory features of front vehicles at different depths. Then random forest is used to select significant kinematic variables which can reflect the potential crash. Finally, a multi-modal deep convolutional neural network (DCNN) combined both video and kinematic data is developed to identify potential collision risk in each 12-second vehicle trip. The analysis results indicate that the proposed multi-modal deep learning model can identify hazardous events within a large volumes of data at an AUC of 0.81, which outperforms the state-of-the-art random forest model and kinematic threshold method.**

I. Introduction

More than 1.25 million people died each year because of road traffic crashes and 90% of the fatalities occurred on the roads in low- and middle-income countries according to World Health Organization [1]. In China, commercial vehicles had attributed to 30.5% of traffic crashes [2], many of them experienced the violations of traffic rules and chaotic driving by pedestrians, bicyclists, and surrounding vehicles, as well as driver distraction. To prevent vehicle crash and understand the accident causation, driving record instruments are required in commercial vehicles. According to the Regulation on the Implementation of the Road Traffic Safety Law in China, the road passenger automobiles, heavy lorry or semi-trailer tractor must be equipped with a driving record instrument. Two types of data are recorded: (1) driving video with a low frame rate and definition, and (2) kinematic data such as velocity and deceleration. Identification of the crash and near-crash events within the data plays an important role in crash and near-crash causal factors assessment. In this work, a high position camera watches father distance from a commercial vehicle captures early dangers in video because of the slow stopping of such vehicles. The camera is installed on the upper side of the windshield to capture far objects with less occlusion. The motion profile samples video frames and stacks them into one image along the time axis so that *spatial-temporal images* are obtained to reflect a long-term traffic conditions. The crowd traffic at distance, relative speed of approaching vehicles, the invasion of other traffic into the lane, etc. are the critical factors causing hard braking later if a driver is not aware of such events. Because of the difficulty in explicit modeling of such scenarios far at front, we employ the *deep learning* method to memorize such "impression" in the driving video. A model trained by multi-modal deep learning method which simultaneously considering vehicle kinematic features and its surrounding traffic environment is proposed in this paper. Three main components include: (1) Motion profile acquisition as temporal-spatial images [7] for the traffic motion, position, and depth of dynamic scenes; (2) A random forest model to analyze the kinematic variable importance and select significant variables; (3) Multi-modal deep convolutional neural network (DCNN) trained with motion profiles and selected kinematic variables. Effective features from the image training data that are related to image trajectory, divergence, density, and Time-to-collision (TTC) are reflected by motion profiles. And the DCNN exploits an efficient co-representation of motion profiles and selected kinematic variables.

The main contributions of this paper are: (1) Driving video information extraction and analysis using the motion profile. (2) A multi-modal deep learning model combining both video and kinematic data. The experiments show that the proposed model outperforms state-of-the-art model (AUC 0.81)

II. Related Works

In recent studies, the combination of kinematic thresholds is used to identify the hazardous event [3]. The sensitivity of this method is 0.62 Jerk, which indicates the differential of acceleration as a widely-used variable in post-accident analysis [4, 5]. The jerk threshold method could achieve 86% accuracy in a dataset with 637 hard-braking events [4]. However, environment factors were also proved to have important impacts on collision according to study [6]. They have not been utilized in these kinematic threshold methods. Not all the hard braking action lead to a hazardous event. It may happen in a crossing road due to the traffic signal. Consequently, efficient method for identifying hazardous driving event using video and kinematics data is in need.

Driving videos, usually processed with computer vision techniques, provide environment factors during a vehicle trip. Works in [7, 8, 9] estimate TTC from motion in driving videos without applying vehicle recognition and depth measuring in

Zhen Gao, Yajun Liu and Ping Sun are with School of Software Engineering, Tongji University, Shanghai 201804 China, (e-mail: gaozhen, liuyajun, pingsun@tongji.edu.cn).

Jiang Yu Zheng is with Department of Computer Science, Indiana University-Purdue University, Indianapolis, IN 46202 USA, (e-mail: jzheng@iupui.edu).

Rongjie Yu, Xuesong Wang are with School of Transportation, Tongji University, Shanghai 201804 China, (e-mail: yurongjie, wangxs@tongji.edu.cn).

This is the author's manuscript of the article published in final edited form as:

Gao, Z., Liu, Y., Zheng, J. Y., Yu, R., Wang, X., & Sun, P. (2018). Predicting Hazardous Driving Events Using Multi-Modal Deep Learning Based on Video Motion Profile and Kinematics Data. 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 3352–3357. https://doi.org/10.1109/ITSC.2018.8569659

prior. Computer vision technique has limited performance in recognizing far road traffic due to a low resolution, and a cut in of other vehicles in complex traffic scenes due to the difficulty in explicit modeling of dynamic environments. In the following, we will discuss the data processing of naturalistic driving video in Section III. Data processing is given in Section IV. The hazardous event identification model including multi-modal DCNN and RF will be introduced in Section V. Experiments will be in Section VI followed with conclusion.

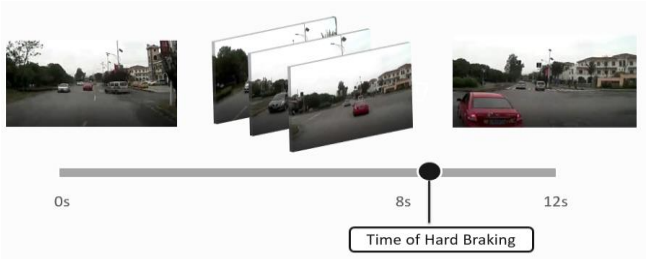


Fig. 1 Timeline of driving event recorded in the video clips for analysis. Each twelve-second vehicle trip contains a hard breaking by driver under the circumstances of crowd traffic, lane occupying by others, fast ego-speed, etc.

III. Naturalistic Driving Video Annotation

The naturalistic driving video and data were obtained from a commercial truck fleet in Shanghai for two years. The obtained videos have 4 frames per second and the resolution is 760×368 pixels. The color of video is turned to low saturation like black-white video in compression. The vehicle velocity were recorded at 1Hz and deceleration at 3Hz. Anytime a high deceleration lower than -0.4g (g: Acceleration of Gravity) occurs, video and data for eight seconds in prior and four seconds after were dumped for machine learning later. Totally, 1959 clips of such video were sampled along with their vehicle control parameters including velocity and deceleration. For all these videos, annotation of dangerous levels and types were carried out by human experts. By observing the selected video clips, we found many critical incidents caused by (1) driver's distraction while front vehicles were approaching quickly; (2) sudden cut-in or U-turn of other vehicles and bicycles into the pathway; (3) some violations of traffic rules by other vehicles causing unpredictable dangers. All these were followed with a sudden braking and/or sharp steering to avoid crash. Depending on the distance where these events happen and the ego-velocity, the severity level is classified to 2 categories in Table I.

Table I. Definition of Severity Level in Driving

Level	***Description***
Hazardous	Any circumstance that requires a crash-avoidance response on the part of any other vehicle [10].
Non-conflict	Any circumstance that affects normal driving and requires driver's reaction. But no conflict objects and potential crash exist [10].

The timeline of recorded driving event is shown in Fig. 1. With a hard braking, most video recorded maneuvers avoided crash, but these should be replaced by a smooth breaking in earlier preparation. To identify whether hazardous event happened during the video clip, video frames which last for 12 seconds are sampled for post-accident analysis.

Many hazardous events happen in a sequential process. They can be observed at a far distance and an early stage, and then become hazardous when they approach to close ranges if the driver did not pay attention. For these reasons, we divide the field of view into three zones to capture frontal vehicles at far, middle, and near ranges respectively as shown in Fig. 2. For the camera obtaining frames as shown in Fig. 2, the horizon is first calibrated at 220 pixel high in the image, which represents the infinity distance. Below the horizon in the image, three zones are selected to cover the ranges of (5, 10], (10, 25], and (25, ∞] meter ahead the vehicle, respectively. The distant zone at high image position observes far traffic while the close one at low position responses to immediate danger of cut in.

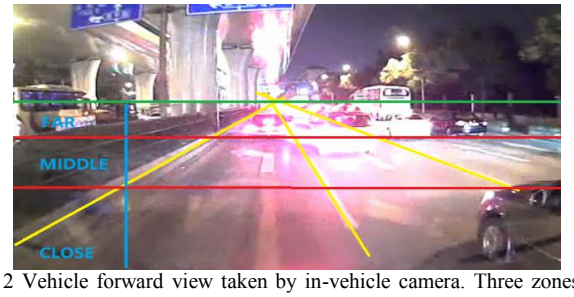


Fig. 2 Vehicle forward view taken by in-vehicle camera. Three zones are located below the horizon to monitor dynamic scenes at three distances on road. They cover ranges in (5, 10), (10, 25), and (25, ∞) meter.

IV. Video Information Extraction and Condensing

To bridge the video signal to the classification of hazardous events and avoid influence caused by complex traffic environment, a data representation that reflects the motion trajectory information more than one video frame is implemented, since the divergence of trajectories can be linked to TTC [9]. Temporal driving video is converted to a spatial-temporal map so that the time, distance, position, and speed of surrounding scenes can be included. This mapping allows the machine learning process to model heterogeneous events. Another merit of it is the data reduction for both big-video learning and on-line real-time event detection during driving. We employ the *motion profile* [7, 8, 9] to record the motion of surrounding traffic.

To grasp the temporal changes at three distances, three motion profiles are generated from three zones. Fig. 3 shows how a motion profile is obtained from a driving video. In details, the color in each zone is vertically averaged to produce a pixel line. For each zone, the lines from all frames are further collected over twelve seconds to form an image called *motion profile* [8]. In the profiles, the object width (size) at corresponding height (depth) and their motion trajectories are recorded as traces. Their density, lateral position, and divergence/convergence rate can be further obtained through computing. The upper zone (far range) has dense and narrow traces from far/small vehicles and background, while lower zone (close range) has uniformed road surface and a bumper if a frontal vehicle gets close. Figure 4(1-3) shows such three motion profiles extracted from a video and their combined image in color is in Fig. 4(4). The data size to process now is the three image slices out of a video volume, which achieves the condensing rate to 3/368, where 368 is the frame height in

pixels. The motion profile obtains the common motion of objects at each range, which is the key factor to cause accident, rather than the identity of objects themselves. The motion profile keeps the important object width and position, rather than object height and shape that is less related to accidents.

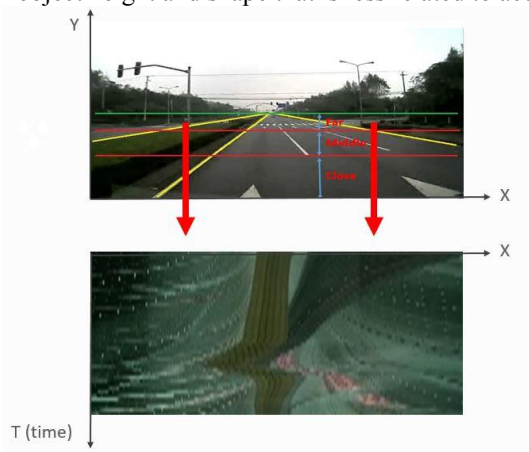
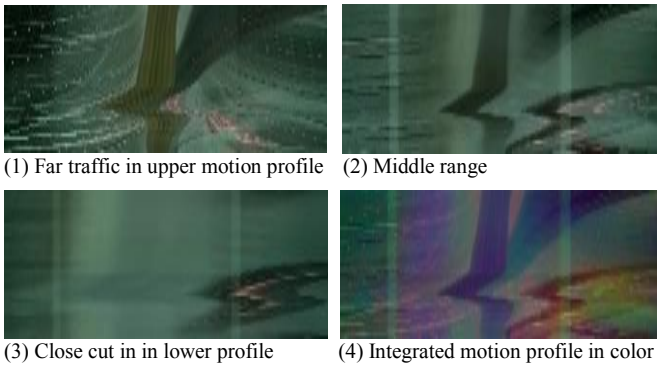


Fig. 3 One example of motion profile obtained from driving videos. Vertical axis of motion profile represents time and horizontal axis represents the x coordinate in video frame.



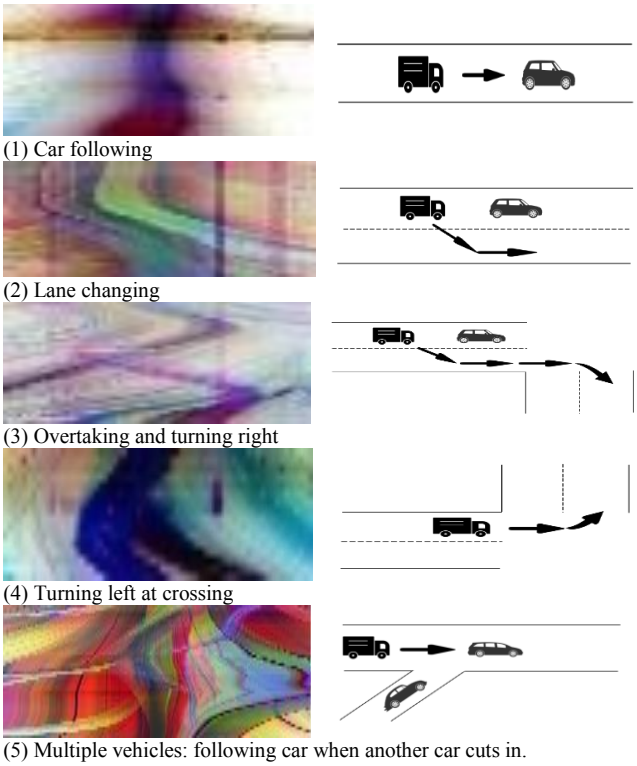
(1) Far traffic in upper motion profile (2) Middle range

(3) Close cut in in lower profile (4) Integrated motion profile in color

Fig. 4 Motion profiles from a video clip of 6 seconds. (1)-(3) are three motion profiles at different image heights. Their integrated display in color is in (4) obtained from gray images of (1)-(3). Middle range vehicle has a trajectory at center. Close range has a target cut-in from right in the view. The common wave across three profiles in the entire x-span comes from the sudden turning of camera/vehicle. The dirty glass at lower profiles draw vertical transparent stripes that bother the traffic flow.

To examine the traffic at different ranges simultaneously and discover their correlation speed in traffic flow, we convert three motion profiles to gray level images (ignore color and illumination factor of traffic), and then combine them into a single color profile, in which the lower profile is set in red channel for alarming, middle profile in green channel for easy observation, and upper profile in blue channel that is less obvious in display. Figure 4(4) shows such a combined motion profile. The motion profile thus converts the temporal information to a spatial representation with a vertical time axis. The object size and position are preserved along the horizontal axis that is the camera angle.

To give more examples, Figure 5 shows five typical driving scenarios represented in motion profiles. In a car following scenario (Fig. 5.1), a main trajectory lies in the middle of the profile, and it gets wider as the time increases. The color of this trajectory is mainly red because the front vehicle was at the close range in video. Figure 5.2 shows a lane changing scenario. Main trajectory is in the middle and it turns right when ego-vehicle is changing lane. Overtaking behavior is in Fig. 5.3 with two trajectories because ego-vehicle overtook one front vehicle and then followed another vehicle at front. Figure 5.4 shows a turning left where a main trajectory is continuous as the ego-vehicle followed the same front vehicle while turning left. Figure 5.5 shows that a vehicle was following a car while another car cut in.



(1) Car following

(2) Lane changing

(3) Overtaking and turning right

(4) Turning left at crossing

(5) Multiple vehicles: following car when another car cuts in.

Fig. 5 Five typical driving scenarios represented by motion profiles.

The key factors to cause forward collision are the density of frontal vehicles at different depths and speed they approach to the camera relatively. The earlier time happening (trace) in the far motion profile (*b* channel) may be less critical in causing a crash, while any happening in the close motion profile (*r* channel) may cause danger immediately. The approach of a vehicle has its size expanding in the profile, i.e., its trajectory diverges [9]. A constantly approaching vehicle from far to close corresponds to a transition of trajectory from high motion profile (*b* channel) to low motion profile (*r* channel), which is a serious case that requires precaution. On the other hand, TTC estimation is unreliable due to the low frame rate (4 fr./sec) of driving videos in the experiments. So deep learning method is adapted to exploit the features of motion profiles.

V. Driving Risk Evaluation Model

A. Variable selection using random forest

The random forest (RF) method is commonly used in many applications involving high-dimensional data [12]. It can be applied for both application and regression. Nominal response is used for classification while numeric response is used for regression. RF can not only obtain predictions but also

identify predictors which are significant. A ranking of predictors that reflects the importance of these variables is available by using RF. This ranking list can be used to select variables with the best predictive ability. Their predictive ability is assessed by VIM (Variable Importance Measure). The formulation of VIM is:

$$\mathrm{VI}M_j^M = \frac{1}{ntree}\sum_{t=1}^{ntree}(MP_{tj} - M_{tj}) \quad (1)$$

Where *ntree* denotes the number of trees in the forest. M_{tj} denotes the error of tree t when predicting all observations that are OOB for tree t before permuting the values of predictor variable X_j, MP_{tj} denotes the error of tree t when predicting all observations that are OOB for tree t after randomly permuting the values of predicting variable X_j.

In this study, an error-based VIM method, also known as MDA (Mean Decrease Accuracy), is adapted to evaluate the predictive ability of 135 variables. These variables consist of speed, acceleration, jerk and their statistical variables, such as mean, variance, maximum, skewness, kurtosis and CV (Coefficient of Variance). Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. An ideal distribution of massive random data should be the normal distribution, the skewness reflects the offset of given data's distribution from normal distribution. We assume that a normal driving event without crash should be in the normal distribution, so skewness shows how abnormal the speed, acceleration and jerk of a specific event are. Similar to skewness, kurtosis is also a measure of the shape of probability distribution. Intuitively, kurtosis reflects the peak value in the mean of a distribution. A high kurtosis of a distribution represents a steep rise or fall. In the driving scenario, a drastic action taken by driver will result in a high kurtosis of the speed/ acceleration/jerk distribution, which means that a potential collision happened in this 12 second trip. The CV, also known as relative standard deviation, is a standardized measure of dispersion of a probability distribution. Different from standard deviation, CV won't be influenced by the data scale. Among all the driving scenarios in our dataset, the mean speed is different in each trip. CV is able to build a uniform measure of the dispersion in both high speed and low speed. The VIM ranking result is shown in Fig. 6.

It shows that accelerationSkew and accelerationKurtosis have the highest MDA, which means they are the most significant variables among the 135 variables. The experiment result of RF proves our assumption above. Driver's drastic response to emergency dose have a significant influence on kinematic variable's distribution. And the statistics which capture the shape characteristics of distribution were finally recommended by the RF model. AccelerationCV, accelerationMin, acceleration8.0s, and speed6s are also significantly more important than other variables. After checking Fig.6, six most significant variables are selected, and they will be fed into the multi-modal DCNN as the kinematic features.

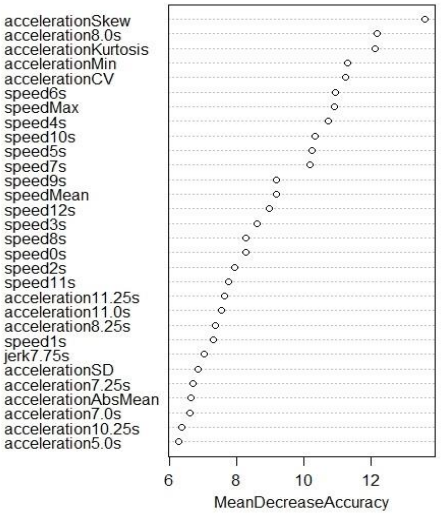


Fig. 6 Variable Importance Measure

B. Multi-modal DCNN Model for Driving Videos

It is difficult to explicitly model the cause of hazardous events in video because of the following reasons: (1) the low video quality (4 frames per second) in color and resolution incapable of measuring the distant objects in shape and speed, (2) the variation of events in video is large across environment, traffic and driver. Therefore, we apply deep learning algorithms to understand the driving videos that caused the potential crash. Deep Convolutional Neural Network (DCNN) [11] is employed to perform the supervised learning. As reported, the CNN can learn object color, local features (edges and blobs), and spatial structure in the image, through convolution and pooling layers in the neural network. This corresponds to our depth, density, and trace position and orientation in the motion profile since the motion profile has converted the temporal motion to spatial layout of traces. The properties of traces for objects and background has been analyzed in [13].

TABLE II. STRUCTURE OF DCNN

No. of layer	***Name***	***Parameters***
1	Input	Image: 227 pixel × 227 pixel × 3 channel Kinematic Variables: 6 dimension
2	Conv1	No. of output = 96, kernel size = 11, stride = 4
4	Pool1	Kernel size = 3, stride = 2
5	Norm1	Local size = 5
6	Conv2	No. of output = 256, kernel size = 5, pad = 2
8	Pool2	Kernel size = 3, stride = 2
9	Norm2	Local size = 5
10	Conv3	No. of output = 384, kernel size = 3, pad = 1
12	Conv4	No. of output = 384, kernel size = 3, pad = 1
14	Conv5	No. of output = 384, kernel size = 3, pad = 1
16	Pool5	Kernel size = 3, stride = 2
17	Full6	No. of unit = 4096 × 1
19	Drop6	Drop rate = 0.5
20	Full7	No. of unit = 4096 × 1
23	Full8	No. of unit = 12 × 1 **(6 units for selected kinematic variables from RF model, other 6 units for images)**
24	Output	2 classes

Since kinematic variables and environment factors both contribute to the identification of non-conflict and hazardous events, multi-modal deep learning model is taken into

consideration. This structure is also inspired by the thought of RGB-D multi-modal DCNN for object recognition [14]. Depth image and RGB image are sampled from different sensors, but the combination of them can improve the recognition ability of deep neural network. It proves that neural network has the potential to exploit the inner-relationship between data from different format and source. In our multi-modal DCNN model, images are processed with convolution and pooling operation, while kinematic features of corresponding images are transferred to the last but two layer of the net without changing any value. The structure and parameters of DCNN is listed in Table II. The input of the network is the motion profile containing both horizontal size and temporal motion and the 6-dimensional kinematic variables. The goal of the network is to identify the trace divergence in the motion profile. To capture sensitive orientation of traces, large filter size at the first layer is specified. In driving videos, the scenes not only depend on the traffic density and size, but also related to weather, illumination, and environment. To avoid overfitting onto small number of samples and learning particular scenes rather than common actions of vehicles, we further invert the color of motion profiles to enhance edge effect (motion trajectory in the motion profile) in the training period. We can observe a dark vehicle shadow against bright road surface in a sunny day when the vehicle is facing the sun, or a white vehicle on a dark asphalt road to have the same motion. We invert the motion profile by

$$P_{inv}(\mathrm{i},\mathrm{j},\mathrm{k}) = 255 - \mathrm{P}(\mathrm{i},\mathrm{j},\mathrm{k}) \quad (2)$$

Fig. 7 shows a contrast between an original motion profile and its inverted one. Trajectory remains the same shape while colors are different.

Fig. 7 Original motion profile and inverted one. The time axis is downward.

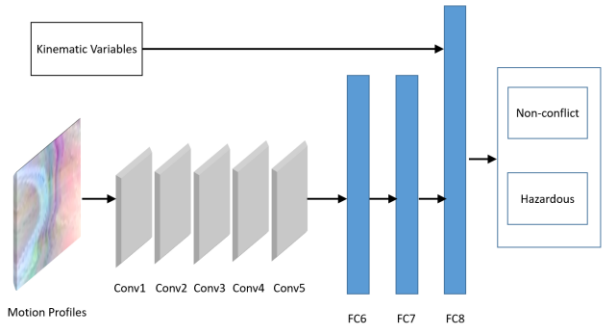


Fig. 8 Architecture of multi-modal DCNN

VI. Experimental Results and Discussion

This section describes the experimental analysis and results. First, a detailed introduction about dataset is stated. Then experiment platform, settings and parameters are introduced. After that, experiment results and comparison with baseline methods are listed. Finally, both quantitative and qualitative analysis about experiment results are presented.

A. Dataset

Four types of collision events on road are considered: (1) forward collision, (2) side collision, (3) T-junction collision, and (4) collision with pedestrians. Among 1959 videos, 954 clips are non-conflict event and 1005 video clips are annotated in hazardous event. Distribution of collision types in the 1005 clips are shown in table III. Because other types of collision except forward collision is relatively limited in sample. Besides, the causation of collision with pedestrian is that people appear at a blind angle of camera, which cannot be reflected by driving videos. So, the major collision types we focused on is forward collision. 1589 events (non-conflict & forward collision) are finally picked up in the experiments and have been divided into training set, validation set and testing set at the ratio of 8:1:1.

TABLE III. DISTRIBUTION OF COLLISION TYPES IN TERMS OF DIRECTION

Collision Type	***Amount***	***Ratio***
Forward collision	644	33%
Side collision	150	8%
T-junction collision	62	3%
Collision with pedestrian	149	7%
Non-conflict	954	49%

B. Experiment Setup

The main experiment setup includes using mini-batch gradient decent for optimization in training multi-modal DCNN. The learning rate is set to 0.01 and the number of maximum training epoch is about 1000. The metric used for representing training loss is the cross entropy. Except for the model proposed in this paper, 4 base-line methods, which are mostly mentioned in recent papers, are applied to the datasets for comparison. All the methods take 80% of the total data as training set, 10% as validation set and 10% as testing set. The implementation and training of multi-modal DCNN is realized by *Tensorflow*. The kinematic variable selection using RF is realized by R. Model training and evaluation is carried out on the workstation with NVIDIA Tesla K40c GPU and Intel Core i7 processors. Training period costs 54,000 steps and 9.2 hours. The accuracy on training dataset is 78%.

C. Experiment Results

Receiver operating characteristic (ROC) curve is plotted to evaluate the identification ability of the proposed model and the baseline methods as shown in Fig. 9. It can be seen that Multi-modal DCNN model has the best performance with the highest area under curve (AUC is 0.81). Threshold with jerk [4] has limited performance on this dataset. It is because that jerk mainly reflects the brake action of drivers. But in the dataset used in this study, not all the braking action finally lead to a hazardous event. Sometimes the drivers braked in advance to keep a safe distance, or a hard braking was done for avoiding running over a red light. RF model using all the kinematic features has the second best performance and its AUC is 0.75. Since environment factors are missing in RF

model, it cannot judge the frontal traffic conditions and that may cause wrong judgement on specific events.

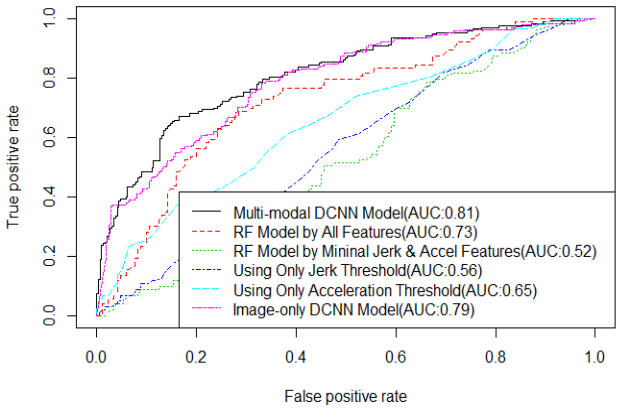


Fig. 9 ROC curve in classifying two cases.

Sensitivity and specificity are used to measure the performance of classification model. These two indexes are computed from

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \qquad \text{specificity} = \frac{TN}{\text{TN+FP}} \tag{3}$$

where TP is True Positive, TN is True Negative, FP is False Positive, FN is False Negative, respectively. *Youden* index is used to compute the best sensitivity and specificity of methods. The formula of computing Youden index is shown as below.

$$J = \text{sensitivity(n)} + \text{specificity(n)} - 1 \tag{4}$$

Where *n* is equal to the number of data points. The threshold with the max Youden index will be selected to compute sensitivity and specificity. And the final result is shown in table IV. It can be concluded that multi-modal DCNN has the best AUC and sensitivity. It proves that combination of kinematic variables and video features can improve the performance of identifying hazardous events. Besides, the jerk threshold method has the best specificity. It is because that jerk threshold can give an accurate judgement on whether a braking action is happened. If no braking action happened during an event, no potential conflict would happen in this event. However, the jerk threshold's sensitivity is the lowest. It can been seen that jerk threshold has limited ability to identify the real hazardous event.

TABLE IV. COMPARISON OF MODELS

Model	***AUC***	***Sensitivity***	***Specificity***
Multi-Modal DCNN	**0.81**	**0.83**	0.67
RF with all features	0.75	0.76	0.66
RF with Jerk and Accel	0.52	0.46	0.63
Jerk Threshold [3,4]	0.56	0.18	**0.89**
Acceleration Threshold [3]	0.65	0.77	0.43

VII. CONCLUSION

The contribution of this paper mainly lies in two aspects. One is the driving video information extraction and analysis using motion profile, which provides the environment factors for analysis of hazardous event's causal factors. The other is a multi-modal deep learning model which combines both video and kinematic data. The experiments show that the proposed model outperforms state-of-the-art models with AUC 0.81. Since driving recording instruments are available in Chinese commercial vehicles, this model can be applied to data reduction of large volumes of driving data without any auxiliary high-precision sensors. In addition, the method to achieve multi-modality of this model also provides a solution to traffic scenario analysis under complex environments. Future research will focus on specifying different crash types and severity levels, which lead to a deeper understanding of road safety relevant events and hazardous event prevention.

ACKNOWLEDGEMENTS

This study was jointly sponsored by Tongji Education Reform Project, Chinese NSF (51522810), Science and Technology Commission of Shanghai Municipality, China (18DZ1200200), Chinese National Engineering Laboratory for Integrated Optimization of Road Traffic and Safety Analysis Technologies, and Chinese 111 Project (B17032).

REFERENCES

[1] World Health Organization. "Road Traffic Injuries." World Health Organization; 2016. Available at: http://www.who.int/mediacentre/factsheets/fs358/en/

[2] Traffic control bureau of the Ministry of Public Security of The People Republic Of China. "The problem of illegal production of part of the truck model highlights the serious hidden danger of commercial vehicle transportation." 2017. Available at : http://www.mps.gov.cn/n2255040/n2255043/c5609936/content.html

[3] P. Miguel A., et al. "Performance of basic kinematic thresholds in the identification of crash and near-crash events within naturalistic driving data." *Accident Analysis & Prevention, vol.* 103, pp. 10-19, 2017

[4] B. Omar. "Assessing safety critical braking events in naturalistic driving studies." *Transportation research part F: traffic psychology and behaviour, vol.* 16, pp: 117-126, 2013

[5] B. Omar, A. Várhelyi. "Development of a method for detecting jerks in safety critical events." *Accident Analysis & Prevention, vol.* 50, pp. 83-91, 2013

[6] J. Wang, et al. "Driving risk assessment using near-crash database through data mining of tree-based model." *Accident Analysis & Prevention, vol.* 84, pp. 54-64, 2015

[7] M. Kilicarslan, J. Zheng, "Temporal video profile from driving video," IEEE Intelligent Vehicle Symposium, pp. 529–551, 2014.

[8] M. Kilicarslan, J. Zheng, "Vehicle collision detection from motion computation in driving video". IEEE Intelligent Vehicle Symposium, 2017.

[9] M. Kilicarslan, J. Zheng, "Predict vehicle collision by TTC from motion using a single camera" IEEE Trans. Intelligent Transportation Systems, pp. 1-12, 2018.

[10] K. Sheila G., et al. "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data." 2006.

[11] K. Alex, I. Sutskever, G. E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. pp. 1097-1105, 2012.

[12] B. Leo. "Random forests." *Machine learning, vol.* 45.1, 5-32, 2001.

[13] A. Jazayeri, H. Cai, J. Zheng, M. Tuceryan, Vehicle detection and tracking based on motion model, IEEE Trans. Intelligent Transportation Systems, vol. 12(2), pp. 583-595, 2011.

[14] A. Wang, et al. "Large-margin multi-modal deep learning for RGB-D object recognition." *IEEE Transactions on Multimedia, vol.* 17.11, pp. 1887-1898, 2015.

[15] Hansen, John HL, et al. "Driver modeling for detection and assessment of driver distraction: Examples from the UTDrive test bed." IEEE Signal Processing Magazine 34.4 (2017): 130-142.

[16] Houenou, Adam, et al. "Vehicle trajectory prediction based on motion model and maneuver recognition." *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013.